

PATENT
450110-04786

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
APPLICATION FOR LETTERS PATENT

TITLE: ADAPTIVE ARTIFICIAL VISION METHOD AND
SYSTEM

INVENTOR: Frederick KAPLAN

William S. Frommer
Registration No. 25,506
FROMMER LAWRENCE & HAUG LLP
745 Fifth Avenue
New York, New York 10151
Tel. (212) 588-0800

Adaptive artificial vision method and system

BACKGROUND OF THE INVENTION

5

1. Field of the invention

The present invention relates to an adaptive artificial vision method and to an adaptive artificial vision system.

10

2. Technical background

In most artificial vision systems, objects are either pre-given or they do not exist.

15

Some systems perform image segmentation based on the particular characteristics of an image (color, boundaries, etc.). These systems have no notion of objects. They just extract regions which seem interesting in themselves. They work well if the background on which the "objects" are presented is known or strongly constrained (e.g. colored objects on a white floor). In such cases the segments automatically extracted can be considered as the "contour" of some objects.

20

Other systems perform object identification given a set of predefined objects that they use as models. If the models are of sufficiently good quality, performances of such systems can be very good. See for example the handbook from S. Ullman entitled "High-level vision: object recognition and visual cognition", MIT Press, Boston, MA, USA, 1996.

25

Unfortunately in some situations, neither of these two conditions can be met. This is particularly true, in the case of robots evolving in natural unknown environment, trying to discover the "objects" present without knowing them in advance. In such cases, segmenting and recognizing objects become a bootstrapping problem that can be summarized in the following way:

30

- Segmentation algorithms do not work well in real-life conditions if no template of the objects is provided.
- Templates of the objects cannot be built without a good segmentation algorithm.

35

This situation leads to a technological deadlock.

Summary of the invention

The present invention aims at overcoming the above-mentioned drawbacks and at enabling efficient bootstrapping of artificial visual recognition even in an unknown environment where objects which are not
5 known in advance are present.

These aims are obtained by an adaptive artificial vision method comprising the following steps:

(a) defining successive couples of synchronized timesteps ($t_{-1}, t ; t, t_{+1} ; \dots$) such that the time difference τ between two synchronized
10 timesteps ($t_{-1}, t ; t, t_{+1} ; \dots$) of a couple of synchronized timesteps is equal to a predetermined time delay τ_0 ,

(b) comparing two successive images ($I_{t-1}, I_t ; I_t, I_{t+1}, \dots$) at each couple of synchronized timesteps ($t_{-1}, t ; t, t_{+1} ; \dots$) spaced by said predetermined time delay τ_0 for obtaining a delta image Δ_t which is the
15 result of the computation of the distance between each pixel of said two successive images ($I_{t-1}, I_t ; I_t, I_{t+1}, \dots$) in view of characterizing movements of objects between said two successive images ($I_{t-1}, I_t ; I_t, I_{t+1}, \dots$),

(c) extracting features from said delta image Δ_t for obtaining a potential dynamic patch P_t which is compared with dynamic patches previously recorded in a first repertory R_d which is progressively constructed
20 in real time from an initial void repertory,

(d) selecting the closest dynamic patch D_i in the first repertory R_d or if no sufficiently close dynamic patch still exists, adding the potential dynamic patch P_t to the first repertory R_d and therefore obtaining and
25 storing a dynamic patch D_i from the comparison of two successive images ($I_{t-1}, I_t ; I_t, I_{t+1}, \dots$) at each couple of synchronized timesteps ($t_{-1}, t ; t, t_{+1} ; \dots$), and

(e) temporally integrating stored dynamic patches D_i of the first repertory R_d in order to detect and store stable sets of active dynamic patches representing a characterization of a reoccurring movement or event
30 which is observed.

When stable sets of active dynamic patches representing a characterization of a reoccurring movement have been detected, the center of the movement is identified and static patches which are at a

In such a case, the extraction of features from the region $R_{x,y}$ may comprise measuring the color change of a pixel compared to its neighbors and computing a color model of the color distribution in the region $R_{x,y}$.

5 Successive couples of synchronized timesteps $(t_{-1}, t ; T+t_{-1} ; T+t ; \dots)$ are separated by a period of time T which is equal to \underline{n} times the predetermined time delay τ_0 , where \underline{n} is an integer which is positive or equal to zero.

However preferably successive couples of synchronized timesteps $(t_{-1}, t ; t, t_{+1} ; \dots)$ are contiguous without any time interruption between two
10 successive couples of synchronized timesteps $(t_{-1}, t ; t, t_{+1})$.

The method according to the invention may further comprise the step of detecting transitions between stable sets of active dynamic patches representing a characterization of reoccurring movements and of constructing transition graphs for predicting complex events comprising a
15 sequence of identified movements.

The invention further relates to an adaptive artificial vision system comprising:

- a clock for defining successive couples of synchronized timesteps $(t_{-1}, t ; t, t_{+1} ; \dots)$ such that the time difference τ between two synchronized
20 timesteps $(t_{-1}, t ; t, t_{+1} ; \dots)$ of a couple of synchronized timesteps is equal to a predetermined time delay τ_0 ,

- inputting means for inputting images $(I_{t-1}, I_t ; I_t, I_{t+1}, \dots)$ provided by a camera at said synchronized timesteps $(t_{-1}, t ; t, t_{+1} ; \dots)$,

- first comparator means for comparing two successive images
25 $(I_{t-1}, I_t ; I_t, I_{t+1}, \dots)$ inputted at each couple of synchronized timesteps $(t_{-1}, t ; t, t_{+1} ; \dots)$ spaced by said predetermined time delay τ_0 for obtaining a delta image Δ_t which is the result of the computation of the distance between each pixel of said two successive images $(I_{t-1}, I_t ; I_t, I_{t+1}, \dots)$,

- first memory means (M_d) for storing dynamic patches D_i
30 representing elementary visual parts for describing characterized movements of objects,

- feature extraction means for extracting features from said delta image Δ_t and producing a potential dynamic patch P_t ,

- second comparator means for comparing a potential dynamic
35 patch P_t which is compared with dynamic patches previously recorded in said first memory means (M_d) ,

- selection means for selecting the closest dynamic patch D_i in the first memory means (M_d) or if no sufficiently close dynamic patch still exists, for recording the potential dynamic patch P_t into the first memory means so that a dynamic patch D_i is stored in the first memory means for each comparison of two successive images ($I_{t-1}, I_t ; I_t, I_{t+1}, \dots$) at each couple of synchronized timesteps ($t_{-1}, t ; t, t_{+1} ; \dots$),

- first temporal integration means comprising computing means for computing during a time T_{F1} corresponding to a predetermined number $N1$ of couples of synchronized timesteps the frequency of each dynamic patch D_i stored in the first memory means and threshold means for defining a set of active dynamic patches comprising dynamic patches D_i whose frequency is higher than a predetermined threshold, and,

- second temporal integration means comprising computing means for computing during a time T_{F2} corresponding to a predetermined number $N2$ of couples of synchronized timesteps the frequency of each set of defined active dynamic patches and threshold means for defining a stable set of dynamic patches corresponding to a reoccurring movement for each set of active dynamic patches whose frequency is higher than a predetermined threshold.

The adaptive artificial vision system further comprises means for identifying the center of a reoccurring movement represented by a stable set of active dynamic patches and means for triggering static pattern recognition for analyzing static patches which are at a predetermined distance d from said center of a reoccurring movement.

BRIEF DESCRIPTION OF THE DRAWINGS

Specific embodiments of the present invention will now be described by way of example only, with reference to the accompanying drawings, in which:

- Figure 1 is a picture illustrating an example of a reoccurring event which may be observed and detected by a system according to the present invention,

- Figure 2 shows those sets of pictures illustrating the construction process of three different dynamic patches contributing to recognizing the event shown on Figure 1,

- Figure 3 is a picture given as an example for illustrating the construction of a set of active static patches during the detection of an event previously recognized by a dynamic patch set,

5 - Figure 4 shows examples of pictures of objects corresponding to static patch sets constructed during the process of event and object recognition,

10 - Figure 5 shows a picture illustrating a reference movement identified by two stable patch sets and three further pictures illustrating further movements identified by three patch sets which are linked to the reference movement,

 - Figure 6 shows the pictures illustrating different movements which may give rise to transitions from one movement to the other and enable the construction of a transition graph, and

15 - Figure 7 is a diagram schematically illustrating an example of architecture of an adaptive artificial vision system according to the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

20 The adaptive artificial vision system according to the invention is conceived in such a manner that it is capable of bootstrapping object and event recognition.

25 The method and system according to the invention start with very crude recognition capabilities. The system may be called impressionist since it is adapted to perceive patches which are elementary visual parts for describing dynamic or static events. A dynamic event is usually defined by a movement whereas a static event is defined by an object.

30 The patches are constructed by the system to describe its visual perception at the lowest level. As the system accumulates visual experiences, it attempts to integrate patches together in order to discover stable sets. This integration happens both in the time and the spatial domain. After a while, the system becomes capable of recognizing reoccurring movements. After reaching this stage, it tries to extract the structure of objects involved in the movements it detects. By these means, it begins to be able to recognize these objects even when they are not
35 moving. Stage after stage, starting from scratch, this artificial system learns to structure its perception into more and more complex representations.

Stable patch sets constitute a higher level representation of things happening and present in the environment (events and objects). Furthermore the way dynamic patch sets are constructed directly influences the way static patch sets are processed. In practice, good dynamic
 5 recognition comes first and helps static recognition to get off the ground.

The process of dynamic recognition will now be described in a more detailed manner with reference to Figures 1, 2 and 7.

A clock 101 defines successive couples of synchronized timesteps such as t_{-1} , t . The time difference τ between two synchronized timesteps, e.g. t_{-1} , t of a couple of synchronized timesteps, is equal to a predetermined
 10 time delay τ_0 .

Successive couples of synchronized timesteps, e.g. t_{-1} , t ; $T+t_{-1}$, $T+t$; ... are separated by a period of time T which is equal to \underline{n} times the predetermined time delay τ_0 , where \underline{n} is an integer which is positive or
 15 equal to zero.

Preferably $n = 0$ and successive couples of synchronized timesteps t_{-1} , t ; t_{+1} ; ... are contiguous without any time interruption between two successive couples of synchronized timesteps.

Images of the environment are obtained by a camera 103 such as a video camera, and inputted to the system via an inputting circuit 102
 20 synchronized by the clock 101. Digital images I_{t-1} , I_t , I_{t+1} , ... of scenes viewed by the camera 103 are therefore inputted by the inputting circuit 102 to a first comparator 104 at defined successive timesteps t_{-1} , t ; t_{+1} ; ...

In the following description it will be assumed that successive
 25 couples of synchronized timesteps t_{-1} , t ; t_{+1} ; ... are contiguous (i.e. $n = 0$ and $T = n\tau_0 = 0$).

A dynamic patch is obtained by comparing two successive images I_{t-1} , I_t ; I_t , I_{t+1} ; ... inputted to the first comparator 104 at each couple of synchronized timesteps t_{-1} , t ; t , t_{+1} ; ... spaced by the predetermined time
 30 delay τ_0 .

A delta image Δ_t constitutes the result of the computation of the distance between each pixel of two successive images I_{t-1} , I_t ; I_t , I_{t+1} , ... in view of characterizing events (movements of objects) which occurred between the two successive images I_{t-1} , I_t ; I_t , I_{t+1} , ...

A filter function such as a threshold function f_{th} may be included in the first comparator 104 in order to keep only the most significant differences.

5 In such a case, $\Delta_t = f_{th} (|| I_t - I_{t-1} ||)$ when a couple of images I_{t-1} , I_t is considered.

A memory M_d is used for storing dynamic patches D_i representing elementary visual parts for describing characterized movements of objects.

The delta image Δ_t is applied to a module 105 which extracts features from the delta image Δ_t and produces a potential dynamic patch P_t .

10 The module 105 for extracting features of the delta image Δ_t may use different known techniques.

For example a gaussian color model of the distribution for each color component may be computed by the module 105.

15 Alternatively, the step of extracting features from the delta image Δ_t in the module 105 may comprise using histograms to model the distribution for color components, shape or texture.

In the case of a color model, a potential dynamic patch P_t represents a color distribution of the delta image Δ_t .

20 The memory M_d comprises a repertory R_d which is progressively constructed in real time from an initial void repertory.

In a module 106, a potential dynamic patch P_t is compared with dynamic patches previously recorded in the repertory R_d of the memory M_d .

25 A selection module 107 permits to select the closest dynamic patch D_i in the memory M_d or, if this is an initial step or if no sufficiently close dynamic patch still exists, the selection module 107 permits to record the potential dynamic patch P_t in the repertory R_d within the memory M_d .

A dynamic patch D_i is therefore stored in the memory M_d for each comparison of two successive images at each couple of synchronized timesteps.

30 On Figure 2, references 20A and 20B illustrate an example of two successive images showing a first movement of a hand with a ball of a certain color (e.g. red). Reference 20C shows the delta image Δ_t corresponding to the successive images 20A, 20B. A dynamic patch D_{20} is constructed from delta image 20C and is stored in the dynamic patch repertory R_d .

35

Similarly, references 22A and 22B illustrate an example of two successive images showing a second movement of a hand with a ball of the same red color and reference 22C shows the delta image corresponding to the images 22A, 22B and enabling the construction of a dynamic patch D_{22} .

5 References 26A, 26B illustrate an example of another couple of two successive images showing a third movement of a hand with a ball of another color, e.g. a yellow ball, and reference 26C shows the delta image corresponding to the images 26A, 26B and enabling the construction of a dynamic patch D_{26} .

10 Dynamic patches D_{20} , D_{22} , D_{26} are all stored in the repertory R_d of the memory M_d .

The module 108 of Figure 7 constitutes first temporal integration means and comprises computing means 108A for computing during a time T_{F1} corresponding to a predetermined number $N1$ of couples of
15 synchronized timesteps the frequency of each dynamic patch D_i stored in the register R_d . Threshold means 108B permit to define a set of active dynamic patches with the dynamic patches D_i whose frequency is higher than a predetermined threshold.

Then the frequency of each patch D_i of the repertory R_d during the
20 last $N1$ couples of timesteps is computed. All the patches D_i which frequency is above a given threshold are said to be active. At a given timestep, there is a set of active patches which are supposed to characterize the event which is experienced by the system.

For example, at the timestep t , with picture 1 of Figure 1, three
25 patches may be considered to be active: $A_t = D_{20}, D_{22}, D_{26}$. The picture 1 of Figure 1 shows a moving hand with a yellow ball and may be deemed to be defined by the dynamic patches D_{20} , D_{22} , D_{26} corresponding to the delta images 20C (first movement of the hand with a red ball), 22C (second movement of the hand with a red ball) and 26C (third movement of the
30 hand with a yellow ball).

The system shown on Figure 7 further comprises a module 109 constituting second temporal integration means.

The module 109 comprises computing means 109A for computing during a time T_{F2} corresponding to a predetermined number $N2$ of couples
35 of synchronized timesteps the frequency of each set A_t of defined active dynamic patches such as D_{20} , D_{22} , D_{26} .

Threshold means 109B permit to define a stable set of dynamic patches corresponding to a reoccurring movement for each set A_t of active dynamic patches whose frequency is higher than a predetermined threshold.

5 Thus the frequency of each active patch is computed and if some active patch set reoccurs sufficiently often, it is considered to be stable. It is in that case assumed that it corresponds to something interesting happening regularly in the environment. An event of higher level than the patches is therefore created. Such a dynamic event, which for example in
10 Figure 1 may be defined as a "showing yellow ball" event, is created and defined by the set of patches involved in the reoccurring movement. In the example of Figure 1, the event is defined by the set of patches D_{20} , D_{22} , D_{26} and these data are stored in memory M_d .

As soon as such a dynamic event is constructed, a special detector is
15 created triggering each time an active patch set corresponds to this event. Figure 2 shows an example of such a detection. The user is showing a yellow ball to the camera. The system characterizes this event with the active patch set $A_t = (D_{20}, D_{22}, D_{26})$. The system recognizes this set as a dynamic event, which has already appeared several times in the past. To
20 have a better insight of how the recognition works, Figure 2 shows for each of the patches involved in the detection the two images that were at the origins of the initial patch creation. D_{20} and D_{22} were created when a red ball was shown to the camera and D_{26} for a similar event involving this time a yellow ball. The two first are characteristics of the special movement of
25 the hand performed when showing a round-shaped object, the last one is more specific about the yellow color involved. The description in patches can be seen as a very simple "language" that the system builds to economically describe what is happening. Every time a new event is perceived, the system recruits, as much as it can, existing patches in its
30 repertoires to account for it.

Thus, if it is assumed that an interesting object can be moved, by opposition with features which are only part of the background, the results obtained by the dynamic recognition process may be used to bootstrap or refocus a process of static recognition.

35 When a dynamic event constituted by a movement has been reliably detected as being recurrent, the system immediately focuses its attention

on the movement center identified in the module 110 of Figure 7 and static pattern recognition achieved by a module 112 is triggered by triggering means 111 so that static patches which are at a predetermined distance d from the center of a reoccurring movement are analyzed to constitute at a given timestep a set of active patches.

On Figure 3 a known event corresponding to a hand movement has been detected (picture 3). In that case, four static patches are considered to be part of it (S16, S29, S34, S40).

The system may compute the frequency of each active static patch set in exactly the same manner as for the dynamic case. If some set is regularly involved in known movements, it is assumed that they correspond to an existing object. A static object is then created. It is defined by the set of static patches regularly perceived.

As soon as the first static object is constructed, the system looks at each timestep for it. It searches the patches detected during the image analysis, looking for the configuration corresponding to the definition of the object. The object, which was first constructed because it was part of a reoccurring movement, can now be detected even if it is not moving. In the example of Figure 1, the system detects both a known movement (hand showing movement) and a known object (the ball).

Figure 4 shows examples of objects corresponding to static patch sets which may be constructed by the system from pictures including the picture 30 of Figure 3. Objects 11 to 14 represent respectively in this example a head, a neck, a ball and a dog shaped robot.

A specific example of static recognition system will be described herebelow:

Static patches are based on salient points in the visual perception I_t . For example, in a human face salient points may be the eyes and the mouth.

These salient points can be determined by, for instance, measuring the color change of a pixel compared to its neighbors. This is typically the method that is used for edge detection.

When a salient point (x,y) is detected, the region $R_{x,y}$ corresponding to the surrounding pixels is defined.

$R_{x,y} = \text{regionaround}(x,y)$

As in the dynamic case, different features may be extracted but according to a specific embodiment, a color model of the pixel of this region is computed. This will constitute a potential static path $S_{x,y}$

$$S_{x,y} = \text{colordistribution}(R_{x,y})$$

This patch is compared to the existing patches in the memory M_d . The closest patch S_i in the memory is selected. If no sufficiently close patch exists, $S_{x,y}$ is added to the memory M_d .

$$S_i = \text{closestpatch}(S_{x,y}, M_s)$$

At each timestep t , the system analyzes the image I_t and produces a set of static patches S_i for describing it. The larger the number of static patches, the better will be the recognition. But as this process goes on at every timestep, one must be careful of not producing an analysis which is too time consuming. Therefore, for performance sake, it can be wise to limit the number of possible static patches. This can be obtained in different ways.

1. By doing the analysis at a lower resolution.
2. By introducing a focus mechanism that only concentrates on a given part of the image for each timestep.
3. By specifying a minimal distance between each salient points.
4. By prefiltering the image, extracting parts that are more likely to contain interesting things (saturated or skin-toned color).
5. By limiting the salient points to a fixed number and only take the most salient ones.

An example of an already recognized event E_1 defined by a stable set of dynamic patches is illustrated on the picture 50 of Figure 5. This example of event E_1 is constituted by a movement of a hand defined by two dynamic patches D_1 and D_4 .

Pictures 51, 52, 53 of Figure 5 illustrate other events E_2 , E_3 , E_4 which are defined by other stable sets of dynamic patches, namely D_1 - D_4 - D_9 , D_1 - D_4 - D_5 and D_1 - D_4 - D_{11} respectively.

Pictures 51, 52, 53 show further movements of a hand with an object constituted by a ball which is held in the hand. Pictures 51 and 53

show very similar movements of hand with balls of different colors. Pictures 51 and 52 show slightly different movements of hand with balls of the same color.

5 The definitions of events E_2 , E_3 , E_4 all include dynamic patches D_1 and D_4 which define event E_1 , and differ from the definition of event E_1 only by the addition of a third dynamic patch (respectively D_9 , D_5 , D_{11}). Thus, similarities may be identified between different events E_1 , E_2 , E_3 , E_4 starting from their structure defined by a stable set of dynamic patches. Events E_2 , E_3 , E_4 may thus be recognized as belonging to the same family of similar events since they are all structurally related to another autonomous event E_1 .
10

Figure 6 illustrates another way of structuring the event space.

If an event E_i defined by a stable set of dynamic patches is recognized at a time step t_1 , the system according to the present invention
15 may observe the new situation created at a further timestep t_2 where an event E_j defined by a stable set of dynamic patches is recognized.

If the time difference between t_2 and t_1 is smaller than a predetermined time interval Δ_t (Δ_t being itself far greater than the predetermined time delay τ_0 between synchronized timesteps t_1 , t , t_{+1} ,...
20 and also greater than the integration time of a stable set of dynamic patches), then a transition is identified between events E_i and E_j and a transition arc is created. Progressively, a transition graph is constructed with transition arcs between events E_i , E_j ,... defining movement types identified by stable sets of dynamic patches.

25 It may be noted that the same movement may be continued during a certain amount of time and a transition arc may exist between the similar event E_i observed at different timesteps t_1 , t_2 . Transition arcs may as well be constructed between different events E_i , E_j observed at successive timesteps t_1 , t_2 .

30 After a certain amount of time where different events have been observed several times, the frequencies of the observed transitions may be used to predict the probability that each identified transition occurs when the same situation is detected.

35 Figure 6 shows three pictures 61, 62, 63 illustrating three different events E_5 , E_6 , E_7 respectively.

In the example of Figure 6, picture 61 shows a hand holding an object from the top and imparting a movement to and fro, this movement constituting the event E_5 defined by a stable set of dynamic patches D_4, D_2 .

Picture 62 shows a hand holding an object from the bottom and
 5 imparting a shaking movement, this movement constituting the event E_6 defined by a stable set of dynamic patches D_3, D_2 .

Picture 63 shows a hand holding an object from the top and imparting a spinning movement, this movement constituting the event E_7 defined by a stable set of dynamic patches D_4, D_3, D_2 .

10 Transition arcs are represented on Figure 6 with the probabilities of each new transition computed from the already observed transitions and the frequencies thereof.

In this example, the probability that an event E_5 follows after itself is 70% whereas the probability that an event E_7 follows after an event E_5 is
 15 30% and the probability that an event E_6 follows after an event E_5 is 10%. The probability that an event E_6 follows after itself is 85% whereas the probability that an event E_5 follows after an event E_6 is 10% and the probability that an event E_7 follows after an event E_6 is 5%. Finally, the probability that an event E_7 follows after itself is 65% whereas the
 20 probability that an event E_5 follows after an event E_7 is 30% and the probability that an event E_6 follows after an event E_7 is 5%.

The system according to the invention does not need to store huge amounts of data since patches or sets of patches are stored in memories M_d, M_s .

25 Moreover in case of need, some prototype patches D_i which are not used in stable sets of patches may be eliminated at regular intervals to reduce the amount of data stored.

It has been checked that the system according to the invention may be implemented in embedded systems without too much processing power.

30 For example, the system may be incorporated in a robot equipped with a 492x362 pixels color CCD camera and connected to a separate computer through a wireless LAN. Such a computer may be for example a SONY VaioNote PCG CIXN running Linux.

The robot is evolving in unknown environments such as a house. In
 35 such conditions, traditional segmentation methods give very poor results, whereas the invention provides a significant improvement. In particular, if

the system is used in an application in which the robot tries to learn the name of objects, a good mechanism for sharing attention about a particular object is a crucial component of the system. In a robot equipped with a system according to the invention, the impressionist vision module is
5 coupled with an attention manager that controls the head of the robot. The robot may thus turn its head towards things that move and things it knows. As the bootstrapping process goes on, the robot displays its progress in recognition by showing a sharper attention behavior.